AD-A102 683 UNIVERSITY OF CENTRAL FLORIDA ORLANDO DEPT OF MATHEM--ETC F/G 12/1
FITTING DISTRIBUTIONS TO DATA, A COMPARISON OF TWO METHODS.(U)
JAN 81 P N SOMERVILLE, S J BEAN F19628-80-C-0004
UNCLASSIFIED AFGL-TR-81-0074 NL

1 OF 1
AD
A102683

END
DATE
FILMED
9 81
DTIC

LEVEL II

12

AD A102683

(18) (19)
AFGL-TR-81-0074

(6)
FITTING DISTRIBUTIONS TO DATA,
A COMPARISON OF TWO METHODS.

(10) Paul N./Somerville
Steven J./Bean

University of Central Florida
Department of Mathematics and Statistics
P.O. Box 25000
Orlando, Florida 32816

(9) Scientific Report No. 2, 1 Oct 79-31 Jan 81,

(11) 30 Jan[redacted]81 (12) 18

Approved for public release; distribution unlimited

(15) F19628-80-C-0004 (16) 6670 (17) 09

AIR FORCE GEOPHYSICS LABORATORY
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
HANSCOM AFB, MASSACHUSETTS 01731

411423

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFGL-TR-81-0074	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Fitting Distributions to Data, A Comparison of Two Methods	5. TYPE OF REPORT & PERIOD COVERED: Scientific Report No. 2, 1 Oct 79 - 31 Jan 1981
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Paul N. Somerville, Steven J. Bean	8. CONTRACT OR GRANT NUMBER(s): F19628-80-C-0004
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Central Florida, Department of Mathematics and Statistics, P.O. Box 25000 Orlando, Florida 32816	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62101F, 667009AF
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Geophisics Laboratory, Hanscom AFB, MA 09731, Contract Monitor/ I.I. Gringorten/LYD	12. REPORT DATE: 30 January 1981
	13. NUMBER OF PAGES: 18
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approval for Public Release, Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Visibility Date, Compacting Data, Curve Fitting, Empirical CDF, Cumulative Distribution Function, Non-Linear Regression, Robust Methods, Least Squares, Maximum Likelihood, Monte Carlo, Contaminated Data

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Monte Carlo methods are used to compare the methods of maximum likelihood and least squares to estimate a cumulative distribution function. When the probabilistic model used is correct or nearly correct, the two methods produce similar results with the MLE usually slightly superior. When an incorrect model is used, or when the data is contaminated, the least squares technique often gives substantially superior results.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

1.0 Introduction

Suppose we wish to estimate $P(X < c)$ where X is a continuous random variable and c is some constant. That is we wish to estimate $F(c)$ where F is the cdf underlying X. Since F is unknown, a model cdf $F(x;\theta)$ is selected, and the vector of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ may be estimated from a random sample. In this situation some care should be taken with respect to the method of estimation of θ. Since we do not know the form of F and we have only postulated the form is $G(x;\theta)$, we should use a method of estimation of θ which is robust. That is, the method by which θ is estimated should result in a value of $G(x;\hat{\theta})$ which is as close as possible to $F(x)$ even when $G(x;\theta)$ is different from $F(x)$.

Also, as any applied statistician knows, data samples often contain some contamination or outliers which are sometimes difficult to detect. The estimation technique used should be robust in the sense that it shoud not be too sensitive to the contamination.

2.0 Methodology

Samples of size N were generated from various underlying distributions in the following manner. First, k random samples of size N were generated from the unit interval with y_{ij} being the i^{th} individual in the j^{th} sample. Using $x_{ij} = F^{-1}(y_{ij})$, k random samples $x_{1j}, x_{2j}, \ldots x_{Nj}$ from $F(x)$ were obtained. We considered four different families of cases as follows:

(i) $F(x)$ is a cdf such as logistic, Weibull, Laplace, etc.

(ii) $F(x)$ is a mixture of two different normal cdf's

(iii) $F(x)$ is a normal cdf with contamination

(iv) $F(x)$ is a Weibull distribution with contamination and without contamination.

We used the normal model for $G(x;\theta)$ in (i), (ii), (iii), and in case (iv) a Weibull model for $G(x;\theta)$ was used. In all of the above cases we estimated the parameters in the model cdf using the maximum likelihood method and a least squares technique.

The least squares estimates were obtained by regressing the model cdf $G(x;\theta)$ on the empirical cdf which in our case requires non-linear regression. The empirical cdf[1] may be defined as below

$$\hat{F}_N(x) = \frac{2i-1}{2N} \qquad \text{for } x_{(i)} \le x < x_{(i+1)} \;,$$

$$= 0 \qquad \text{for } x < x_{(1)} \;,$$

where $x_{(i)}$ is the i^{th} order statistic. Now, the vector of parameters θ is estimated by selecting those values such that

$$\sum_{i=1}^{N} \left(G(x_{(i)};\theta) - \frac{2i-1}{2N} \right)^2$$

is minimized.

Since the above minimization does not usually yield a linear system of normal equations, the parameters must be estimated using non-linear techniques. We used the linearization or Taylor series method which is described in Draper and Smith (1966). There are a number of other possibly more efficient methods. However the linearization method gave us good results with respect to computer time, and it was easily programmed in SAS MATRIX. Most all non-linear techniques

[1] Often i/N is used in place of $(2i-1)/N$ in the definition of $\hat{F}_N(x)$. A more general form is $(i-c)/(n-2c+1)$, where the value of c depends on the distribution. The values of 0 and 3/8 are then used for the uniform and normal distributions, respectively. For further details see Hahn and Shapiro (1967).

require initial values to estimate the parameters. One method used when $\theta = (\theta_1, \theta_2)$ was as follows:

(i) Let $x_{(i)}$ be the i^{th} order statistic such that i/N is between .15 and .25, and let $x_{(j)}$ be the j^{th} order statistic such that j/N is between .75 and .85. That is, select two order statistics, one in the lower and one in the upper tail.

(ii) let $G(x_{(i)};\theta_1,\theta_2) = i/N$ and

$$G(x_{(j)};\theta_1,\theta_2) = j/N$$

and solve the system for θ_1 and θ_2.

This works in some cases such as the case when $G(x;\theta)$ is the Weibull distribution.

The maximum likelihood estimate of θ is the value of θ which maximizes the likelihood function L. If the probability density function for the distribution is given by $g(x,\theta)$, and $x_1, x_2, \ldots, x_N$ are N values chosen randomly from the distribution, then L is given by

$$L = g(x_1,\theta)g(x_2,\theta)\ldots g(x_N,\theta) \quad .$$

The maximum likelihood estimate of θ is usually obtained by differentiating L with respect to θ, setting the resulting derivative equal to zero and solving for the value of θ. For most distributions, the solution is not straight forward and is best obtained by the use of an iterative technique.

3.0 Comparison of Maximum Likelihood and Least Squares

Local and global errors were calculated as a means to compare the two estimation techniques with respect to the underlying cdf. Let C_p be the value such that $F(C_p) = P$. The local error for a specified value of P for the j^{th} sample is given by

$$d_{pj} = G(C_p;\hat{\theta}_j) - P.$$

We define errors E_1 (local), E_2 and E_3 (global) as follows:

$$E_1(p) = \sum_{j=1}^{k} |d_{pj}|/k \quad \text{(The average local error at P over k samples)}$$

$$E_2 = \sum_{j=1}^{k} (\underset{p}{\text{Max}}|d_{pj}|)/k$$

$$E_3 = \sum_{j=1}^{k} \sum_{\text{all } p} |d_{pj}| /Lk$$

where L is the number of values of p used.

For E_2 and E_3 the max and sum were taken over the grid {.01, .05, .1, .2, .3, .4, .5, .6, .7, .8, .9, .95, .99}, and L, the cardinality of this set is 13. In all of the results reported in this paper k = 75. That is, the results are based on 75 random samples of size N. N = 25 in all of the reported results. However, we did use samples of size 10 and 75, and observed the same general pattern as that for samples of size 25.

3.1 Normal Model Used on Various Distributions

Table 3.1 compares the method of maximum likelihood and least squares when the data is sampled (generated) from each of several different distributions and the model used is the normal distribution. When the sample comes from the normal, Laplace, rectangular, symmetric triangular and the Cauchy distributions, the results are independent of the values of the parameters of the distribution.

This is not the case for the Weibull and gamma. The forms we used are given by

$$F(x;\alpha,\beta) = 1 - e^{-\alpha x^{\beta}} \qquad \text{(Weibull)}$$

$$F(x;\eta,\lambda) = (\lambda^{\eta} / \Gamma(\eta)) \int_0^x t^{\eta-1} e^{-\lambda t}\, dt \quad \text{(gamma)}.$$

It is to be expected that for a normal distribution and a normal model, the MLE would be superior. However the difference is small. For the logistic distribution the two methods differ little. For the rectangular and triangular distributions, the MLE is superior, while for the Laplace and the Cauchy, the LSE is superior. The superiority of the LSE is substantial in the case of the Cauchy distribution.

The Weibull is well approximated by the normal when β is between 2 and 6, as is the gamma distribution for large values of η (the skewness and kurtosis of the gamma distribution are given by $2/\sqrt{\eta}$ and $3 + 6/\eta$ respectively). Because of this it is not surprising that the MLE out-performs the LSE for the Weibull (1,4) and gamma (9,2) and the LSE is superior for Weibull (1,1), gamma (3,3) and gamma (2,1).

The difference between the two methods when the Cauchy distribution is the true underlying distribution is illustrated in Figure 3.1. Figure 3.1 shows the true cdf, (the Cauchy distribution), the empirical cdf, and the two normal model cdf's using the MLE and LSE for μ and σ for a typical sample. This graph shows the normal/LSE model giving a much closer approximation to the Cauchy than the normal/MLE model.

3.2 Normal Model Used on the Mixture of Two Normals

Table 3.2 gives the results for the MLE and LSE when the data is from a mixture of two different normal distributions. We use (1-v) and v to denote the respective weights of N(0,1) and N(a,b) distributions.

When the means and variances of the two normal distributions differ but little (that is when a is not far from zero and b is not far from 1), there is little difference between MLE and LSE. The LSE offers substantial improvement when the two mixing normals differ more. Figure 3.2 illustrates this with the graph of the mixture of two normals (.8 N(0,1) and .2 N(3,9)), the empirical cdf, the normal/MLE, and the normal/LSE.

3.3 Normal Model Used on a Normal Distribution with Contamination

In this case the true underlying distribution is normal. However, some of the data has been contaminated or altered in some way. This frequently happens in actual case studies, and sometimes it is difficult to recognize the altered information. This is different from the previous case in that the true underlying distribution discussed in Section 3.2 is an actual mixing of two distinct normals, and in this case the true underlying distribution is a single normal distribution.

Table 3.3 gives the results for the two methods. The underlying distribution is N(0,1) with a proportion v of the data altered. The altered observations were assumed to be N(a,b). In practice the contamination may take other forms, but the given alteration serves to illustrate the effects of contamination on the estimation methods. It seems clear that the method of least squares gives much better results than maximum likelihood even when only modest contamination is present.

Figure 3.3 illustrates the difference between the two methods when a N(0,1) distribution is contaminated (4%) with a N(0,9).

It should be noted that the LS model appears to be very stable under contamination while the ML model seems quite sensitive to contamination.

3.4 Weibull Model Used on Weibull Data With and Without Contamination

The samples were generated from a Weibull (2,6) population with a proportion v of the data being "contaminated." The contamination was effected by the transformation $\sqrt{b}\,x + a$. Table 3.4 compares the MLE and LSE for various proportions of contamination and combinations of a and b. When v = 0, i.e., there is no contamination, the MLE is slightly superior. For the three cases where there is contamination, the LSE shows a definite superiority.

F(x)		E_1 (Local Errors) P .05	.5	.95	E_2	E_3
Normal	ML	.023	.070	.028	.0861	.0452
	LS	.027	.071	.031	.0926	.0479
Laplace	ML	.035	.074	.030	.129[illegible]	.0574
	LS	.036	.071	.032	.1148	.0551
Rectangular	ML	.024	.076	.026	.1169	.0529
	LS	.041	.084	.046	.1235-	.0602
Triangular (Symmetric)	ML	.024	.077	.021	.0900	.0465+
	LS	.030	.081	.028	.1001	.0519
Logistic	ML	.029	.078	.030	.1024	.0511
	LS	.029	.080	.028	.1031	.0510
Cauchy	ML	.139	.077	.166	.2896	.1330
	LS	.049	.077	.048	.1279	.0606
Weibull (1,4)	ML	.025	.063	.025	.0799	.0418
	LS	.028	.070	.032	.0913	.0476
Weibull (1,1)	ML	.104	.126	.036	.1742	.0834
	LS	.094	.093	.044	.1618	.0739
Gamma (9,2)	ML	.030	.071	.027	.0963	.0471
	LS	.034	.073	.031	.1011	.0497
Gamma (3,3)	ML	.047	.087	.031	.1184	.0565-
	LS	.049	.078	.036	.1160	.0557
Gamma (2,1)	ML	.064	.099	.033	.1337	.0638
	LS	.049	.084	.041	.1287	.0597

TABLE 3.1

A Comparison of Maximum Likelihood and Least Squares

Using the Normal Model

v	a	b		E_1 (Local Error) P .05	.5	.95	E_2	E_3
.08	3	1	ML LS	.033 .027	.067 .073	.026 .035	.0946 .0975	.0458 .0479
.20	3	1	ML LS	.035 .035	.080 .062	.023 .034	.1048 .1011	.0459 .0450
.52	3	1	ML LS	.020 .041	.058 .034	.027 .051	.0917 .0871	.0386 .0414
.08	3	9	ML LS	.050 .028	.071 .074	.042 .032	.1217 .0980	.0601 .0489
.20	3	9	ML LS	.071 .032	.085 .067	.031 .047	.1555 .1103	.0700 .0495
.52	3	9	ML LS	.065 .069	.110 .072	.046 .033	.1441 .1284	.0657 .0554
.52	0	9	ML LS	.055 .031	.067 .069	.057 .037	.1132 .0956	.0608 .0498
.52	0	1/9	ML LS	.028 .036	.071 .072	.033 .044	.1260 .1044	.0536 .0516

TABLE 3.2

A Comparison Between Maximum Likelihood and Least Squares Using the Normal Model on Data from the Mixture of Two Normals

v	a	b		E_1 (Local Errors) P .05	.5	.95	E_2	E_3
.08	2	1	ML LS	.025 .028	.084 .079	.052 .047	.1066 .1044	.0574 .0565
.20	2	1	ML LS	.021 .025	.129 .116	.117 .100	.1729 .1562	.0950 .0861
.08	3	1	ML LS	.031 .027	.089 .080	.087 .056	.1329 .1093	.0719 .0592
.20	3	1	ML LS	.027 .028	.155 .122	.203 .145	.2436 .1883	.1292 .1009
.08	3	9	ML LS	.051 .030	.082 .075	.107 .047	.1485 .1024	.0800 .0549
.20	0	9	ML LS	.102 .046	.064 .069	.102 .054	.1583 .1096	.0879 .0587

TABLE 3.3

A Comparison of Maximum Likelihood and Least Squares

Using the Normal Model on Contaminated Normal Data

v	a	b		E_1 (Local Errors)			E_2	E_3
				p				
				.05	.5	.95		
.00			ML	.022	.067	.025	.083	.043
			LS	.024	.070	.032	.092	.046
.08	1	1	ML	.102	.063	.262	.262	.135
			LS	.023	.073	.057	.104	.056
.08	.5	2	ML	.085	.062	.231	.233	.120
			LS	.023	.073	.057	.104	.056
.08	0	4	ML	.086	.059	.225	.229	.118
			LS	.023	.073	.055	.103	.055

TABLE 3.4

A Comparison of Maximum Likelihood and Least Squares

Using the Weibull Model on Data

from a Weibull (2,6) with 100 v % Contamination

4.0 An Example of Fitting a Weibull Model to Visibility Data Using Maximum Likelihood and Least Squares Techniques

Table 4.1 gives the empirical cdf, the Weibull/LSE fit, and the Weibull/MLE fit for visibility data at 10 a.m. in February at Mildenhall, England. We were concerned with the estimation of $P(X \leq x)$ where x is any positive real number, and X is visibility in miles. The data was the result of approximately ten years of observations. The object was to produce a simple formula from which the probability of visibility events could be quickly evaluated. Such a formula would "compact" the data, and be useful for simulation models.

The Weibull model had previously been used for a number of other locations for various times of day and year. For the data in the table, the MLE and LSE give very similar results. Because of its robust properties, we have preferred the results from the LSE.

X MILES	0	¼	5/16	½	5/8	3/4	1	1¼	1½	2	2½	3	4	5	6
OBSERVED FREQUENCY	.000	.031	.034	.047	.065	.081	.113	.152	.180	.247	.343	.392	.453	.557	.6[illegible]
LSE FIT	.000	.027	.035	.059	.075	.091	.124	.156	.188	.251	.310	.366	.467	.555	.6[illegible]
MLE FIT	.000	.027	.035	.059	.074	.090	.123	.154	.186	.247	.305	.359	.459	.545	.61[illegible]

TABLE 4.1: The Empirical C.D.F., Weibull/LSE Fit, and the Weibull/MLE Fit for Visibility at Mildenhall, England, February, 10 a.m.

5.0 Summary and Conclusions

The method of maximum likelihood and a least squares technique have been compared under a variety of different situations when the purpose of the estimation was to estimate the cdf. When the probabilistic model used was correct or nearly correct the two methods produced very similar results with the MLE usually slightly better. However, when the model used was wrong or the data was contaminated, the least squares technique often gave substantially better results.

Thus, it appears that the LSE are most useful when the underlying probability distribution is not clearly established or when the sample information has possible outliers. In these situations the LS model exhibits a great deal more stability or robustness than the ML model.

The maximum likelihood method is frequently the only method used for parameter estimation. Our results are in agreement with the statements of Berkson (1980) and LeCam (1980) which point out that maximum likelihood procedures should not be used exclusively without regard to the purposes for which the estimates are required.

6.0 References

Berkson, Joseph, 1980. Minimum Chi-Square, Not Maximum Likelihood! The Annals of Statistics 8, 457-469.

Draper, N.R. and H. Smith, 1966. Applied Regression Analysis, Wiley, New York.

LeCam, 1980, Discussion of "Berkson, Joseph., 1980, Minimum Chi Square, Not Maximum Likelihood!" The Annals of Statistics 8, 473-478.

Hahn, Gerald J. and S.S. Shapiro, 1967. Statistical Models in Engineering, Wiley, New York, p. 293.

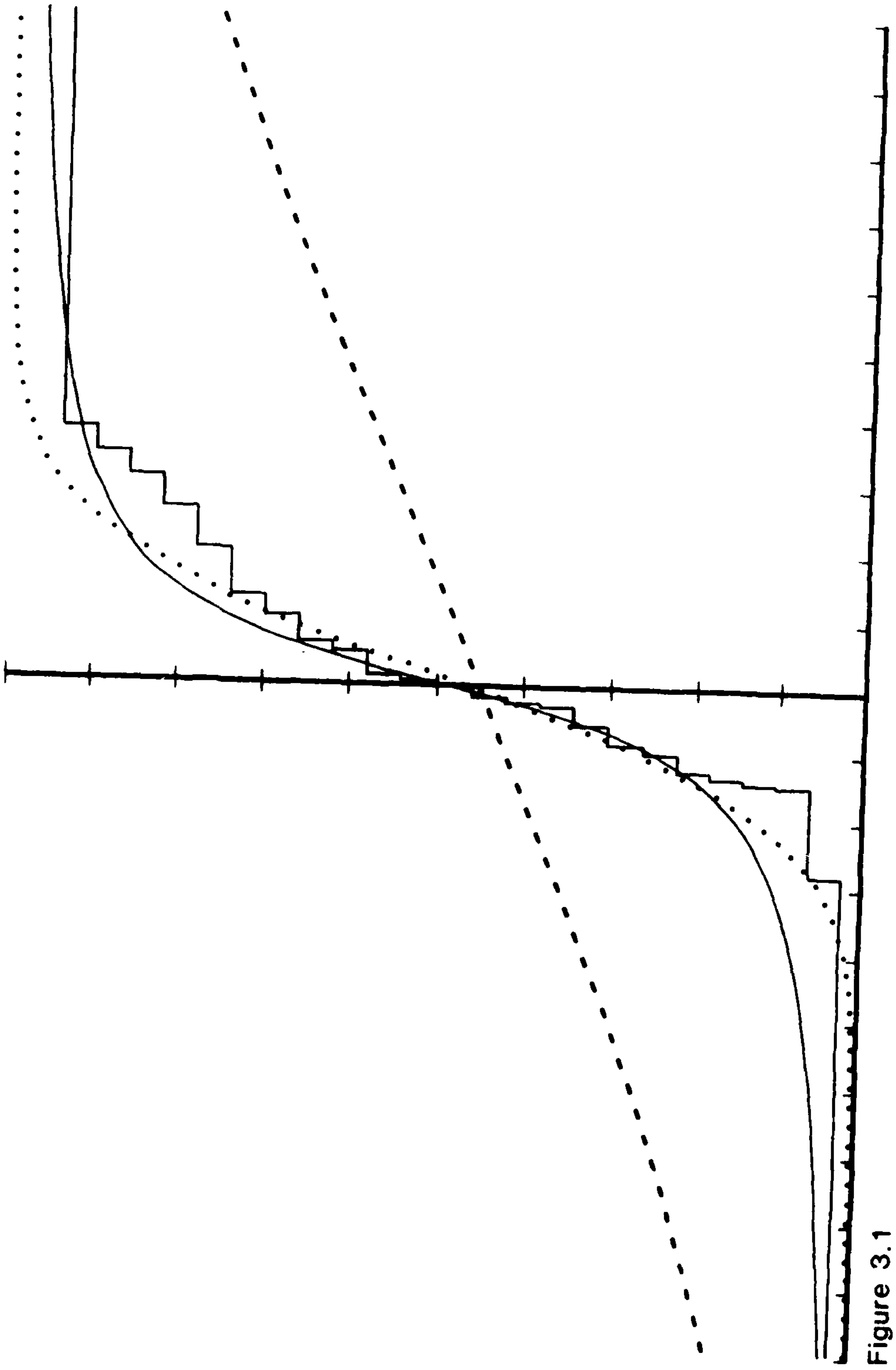

Figure 3.1

CAUCHY DISTRIBUTION

True Cauchy Curve = Solid line, MLE Normal Curve = - - - -

LSE Normal Curve = ...

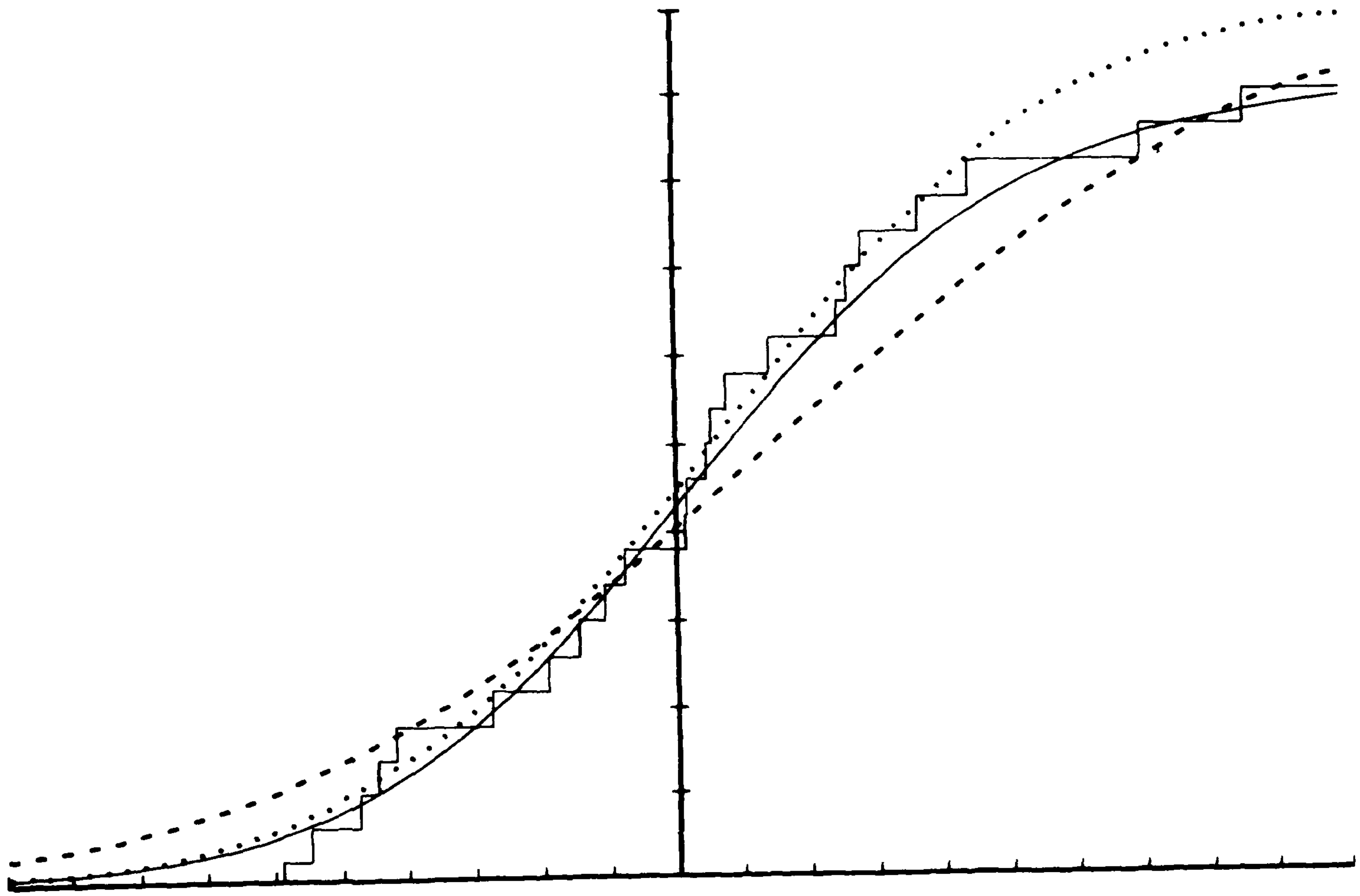

Figure 3.2

Mixture of two NORMAL DISTRIBUTIONS
80 % N(0,1) and 20 % N(3,9) = Solid curve
MLE Normal Curve = ---, LSE Normal curve = ···

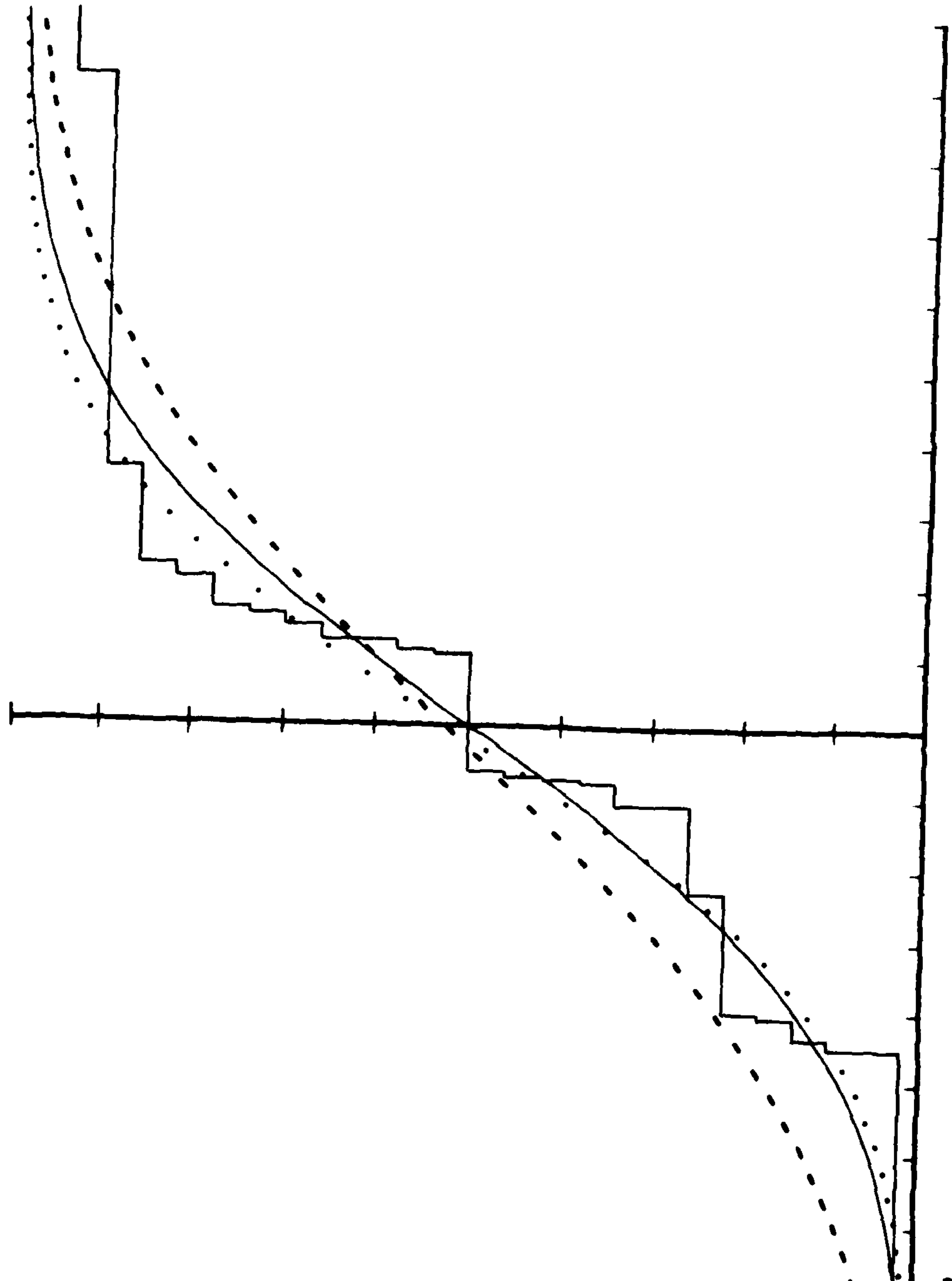

Figure 3.3 N(0,1) with 4% contamination of N(0,9)
N(0,1) = Solid Curve
MLE Normal Curve= - - - ,LSE Normal Curve= . . . .

END

DATE
FILMED
9-81

DTIC